THEORY OF COMPUTING, Volume 14 (20), 2018, pp. 1–29
www.theoryofcomputing.org

Simplified Separation of Information and Communication

Anup Rao* Makrand Sinha†

Received September 20, 2016; Revised January 2, 2018; Published December 27, 2018

Abstract: We give an example of a Boolean function whose information complexity is exponentially smaller than its communication complexity. Such a separation was first demonstrated by Ganor, Kol and Raz (J. ACM 2016). We give a simpler proof of the same result. In the course of this simplification, we make several new contributions: we introduce a new communication lower-bound technique, the notion of a *fooling distribution*, which allows us to separate information and communication complexity, and we also give a more direct proof of the information complexity upper bound.

We also prove a generalization of Shearer's Lemma that may be of independent interest. A version of Shearer's original lemma bounds the expected mutual information of a subset of random variables with another random variable, when the subset is chosen independently of all the random variables that are involved. Our generalization allows some dependence between the random subset and the random variables involved, and still gives us similar bounds with an appropriate error term.

ACM Classification: F.1.1, F.1.2, F.2.3

AMS Classification: 68Q05, 68Q10, 68Q17, 68Q85, 68Q87

Key words and phrases: communication complexity, information complexity, fooling distribution, separation of complexity measures

A preliminary version of this paper appeared in ECCC as Technical Report TR15-057.

*Supported by an Alfred P. Sloan Fellowship, the National Science Foundation under agreement CCF-1016565, an NSF Career award, and by the Binational Science Foundation under agreement 2010089.

†Supported by the National Science Foundation under agreement CCF-1016565, and by the Binational Science Foundation under agreement 2010089.

 DOI: 10.4086/toc.2018.v014a020

1 Introduction

A fundamental question in the study of communication complexity is whether the information complexity of a communication problem is the same as its communication complexity. If the messages in a protocol reveal a small amount of information, does that mean that the protocol can be simulated using few bits of communication? When the protocol is deterministic and one-way, a precise answer was given by Shannon [26] who defined the notion of *entropy*, $\mathbf{H}(M)$, to measure the information content of a message M, and showed that the number of bits of communication can always be made at most $\mathbf{H}(M)+1$ in expectation, which is tight.

For randomized and interactive (multi-round) protocols, this question was made explicit in a sequence of papers. Chakrabarti, Shi, Wirth and Yao [8] defined what we now call the *external information cost* of a protocol, which measures the information learned about the inputs by an external observer of the messages. If M denotes the messages, R denotes the shared randomness and X,Y the inputs, the external information cost is defined to be the mutual information $\mathbf{I}(XY : M \mid R)$. Barak, Braverman, Chen and Rao [2] defined the *internal information cost* of a protocol as $\mathbf{I}(X : M \mid YR) + \mathbf{I}(Y : M \mid XR)$, the information learned by the parties about the input to the other party. The internal information cost is never greater than the external information cost.

For bounded-round protocols, Harsha, Jain, McAllester and Radhakrishnan [16] (see also [4]) showed how to give optimal simulations in terms of the external information cost, and Braverman and Rao [5] gave near-optimal simulations in terms of the internal information cost (up to a $1+o(1)$ factor). However, many of the best known simulations for interactive protocols are not known to be optimal. We know how to simulate any interactive protocol with external information cost I and communication C using a protocol with communication $O(I\cdot\log^2 C)$ [2]. We also know how to simulate a protocol with internal information I and communication C using a protocol with communication $O(\sqrt{IC}\cdot\log C)$ [2]. Braverman [3] showed how to simulate any protocol with internal information cost I using communication $2^{O(I)}$. Natarajan Ramamoorthy and Rao [21] gave better simulations when one party reveals less information than the other. Very recently, Sherstov [27], building on the work of Kol [19], showed that when the input distribution is a product, protocols with internal information I can be simulated with communication $O(I\cdot\log^2 I)$, which is almost optimal.

These results are closely tied to communication lower bounds. Information theory based methods for proving lower bounds on the communication complexity of disjointness [17, 25, 1] can be seen as precursors to some of them. They have been used to give answers to long-standing questions like the direct sum [2] and direct product [6] questions in communication complexity.

Braverman and Weinstein [7] (see also [18]) showed that any Boolean function $f(x,y)$ that can be computed with internal information cost I must have a (nearly) monochromatic rectangle (i. e., a subset $R = S\times T$ of the inputs where the function is essentially constant) of density $2^{-O(I)}$, and so large discrepancy. This means that upper bounds on the size of monochromatic rectangles cannot be used to prove lower bounds on the communication complexity of functions that have small information complexity.[1] So, for a long time, all known methods for proving lower bounds for communication failed to prove lower bounds on functions that have large (0 and 1) monochromatic rectangles. This pointed to a

[1]The information-based methods for proving lower bounds on disjointness also prove that disjointness does not have large 1-monochromatic rectangles.

significant weakness in our ability to prove new communication lower bounds, since certainly, there are functions that require large communication and still have large monochromatic rectangles. One can plant large monochromatic rectangles into a random function to obtain such an example with high probability.

In a remarkable sequence of papers, Ganor, Kol and Raz [13, 15] showed that there is a function with internal information cost I with respect to a distribution that requires $2^{\Omega(I)}$ communication under the same distribution. This proved that Braverman's simulation [3] is tight. Their proof gives a method to prove communication lower bounds on functions that have many large monochromatic rectangles, potentially leading to fundamentally different methods to prove lower bounds on communication problems.

In this paper, we build on the work of [15] and give a new proof of their main result. Our proofs are shorter and we find them more intuitive. For parameters $k,n \in \mathbb{N}$, we define a Boolean function called the *k-ary pointer jumping function* with inputs X,F (given to Alice) and Y,G (given to Bob) and a distribution $q(X,F,Y,G)$ on its inputs. We show that there is a protocol with small internal information cost that computes the k-ary pointer jumping function on the distribution $q(X,F,Y,G)$:

Theorem 1.1. *There is a randomized protocol for the k-ary pointer jumping problem under the input distribution $q(X,F,Y,G)$ that has internal information cost $O((\log k+\log\log n)\cdot 2^{(2\log n)/k})$ and errs with probability at most $4/\log n$.*

On the other hand, we show that no protocol with communication complexity significantly smaller than $\min\{k,\log n\}$ can compute the same function on the distribution $q(X,F,Y,G)$.

Theorem 1.2. *For large enough values of k and n, every protocol for the k-ary pointer jumping function that has communication complexity at most $\varepsilon^3\cdot\min\{k,\log n\}$ errs with probability at least $1/2-8\varepsilon$ on inputs drawn from the distribution $q(X,F,Y,G)$.*

Note that the statement above applies to all protocols, whether they be deterministic or use public and private randomness. Setting $n=2^k$ in Theorem 1.1 and Theorem 1.2, we obtain our main result: when the inputs are sampled from the distribution $q(X,F,Y,G)$, the internal information complexity of the k-ary pointer jumping function is $O(\log k)$ but the communication complexity is $\Omega(k)$.

Corollary 1.3. *There is a randomized protocol for the k-ary pointer jumping function under the input distribution $q(X,F,Y,G)$ that has internal information cost $O(\log k)$ and errs with probability at most $4/k$. Moreover, for large enough k, every protocol for the k-ary pointer jumping function that has communication complexity at most $\varepsilon^3 k$, errs with probability at least $1/2-8\varepsilon$ on the input distribution $q(X,F,Y,G)$.*

One of the new ideas that simplify our proofs is the use of a new technique, the notion of a fooling distribution to prove communication lower bounds. This gives us another method to separate information and communication apart from the *relative discrepancy* method introduced in [15]. We describe this technique in Section 3.3 and compare it with the relative discrepancy method in Section 4.3.

We remark that our function and distribution is a simpler variant of the *bursting noise function* defined in [15] and as such we recover the same parameters in our main theorems as in the work of [15]. For example, [15] proves that any protocol with communication at most 2^t computing the bursting noise function with parameter $t\in\mathbb{N}$ must have error at least $1/2-2^{-t}$. We recover the same parameters from Corollary 1.3 by setting $k=512\cdot 2^{4t}$ and $\varepsilon=2^{-t-3}$. An analogous statement also holds for the information cost upper bound in Corollary 1.3. Moreover, the inputs to the k-ary pointer jumping function are of length N where $\log\log\log N=\Theta(k)$ so the communication and information complexity of this function are really small in terms of the input length (a similar statement holds for the bursting noise function).

1.1 Closely related work

Our work and the work of [15] proves that there is a Boolean function which has widely different information complexity and communication complexity under a carefully designed input distribution. One can also ask if information complexity and communication complexity are exponentially separated for the worse-case input distribution (the non-distributional setting as defined in [3]). Fontes, Jain, Kerenidis, Laplante, Lauriére and Roland [12] showed that the *relative discrepancy* technique introduced in [15] cannot be used to separate information and communication complexity for a Boolean function in the non-distributional setting.

After the dissemination of our paper, Ganor, Kol and Raz [14] gave an exponential separation of external information and communication for a *relation* (search problem) that holds in the non-distributional setting. The new proof uses a clever reduction to the randomized communication complexity of set disjointness. However, as the new result of [14] proves a separation only for a search problem, our ideas still give the most direct proof separating information complexity and communication complexity for Boolean functions.

In another closely related work, Natarajan Ramamoorthy and the second author [22] used one of the technical lemmas (Lemma 4.1) clarified in this paper to give a simpler proof showing that the randomized communication complexity of the Greater-Than function on n bits is $\Omega(\log n)$. The same lower bound was previously proved by Viola [28] and also by Braverman and Weinstein [7] using different techniques.

1.2 Organization

We start with the preliminaries in Section 2. Following this in Section 3, we define the k-ary pointer jumping function and our results in a little more detail. We prove the communication lower bound in Section 4. In Section 5, we bound the information complexity of the k-ary pointer jumping problem.

2 Preliminaries

2.1 Probability spaces and variables

Throughout this paper, log (and ln) denotes the logarithm taken in base two (and base e). We use $[k]$ to denote the set $\{1,2,\ldots,k\}$ and $[k]^{<n}$ to denote the set of all strings of length less than n over the alphabet $[k]$, including the empty string. The notation $|z|$ denotes the length of the string z.

Random variables are denoted by capital letters (e. g., A) and values they attain are denoted by lower-case letters (e. g., a). We say a random variable A determines another random variable B if given the value of A, the value of B is fixed. Events in a probability space will be denoted by calligraphic letters (e. g., $\mathcal{E}$). Given $a=(a_1,a_2,\ldots,a_n)$, we write $a_{\leq i}$ to denote $a_1,\ldots,a_i$. We define $a_{<i}$ similarly. We write a_S to denote the projection of a to the coordinates specified in the set $S \subseteq [n]$.

Given a probability space p and a random variable A in the underlying sample space, we use the notation $p(A)$ to denote the probability distribution of the variable A in the probability space p. We will often consider multiple probability spaces with the same underlying sample space, so for example $p(A)$ and $q(A)$ will denote the distribution of the random variable A under the probability spaces p and q respectively with the underlying sample space of p and q being the same. We write $p(A \mid b)$ to denote the distribution of A conditioned on the event $B = b$. We write $p(a)$ to denote the number $\mathbb{P}_p[A = a]$ and $p(a \mid b)$ to denote the number $\mathbb{P}_p[A = a \mid B = b]$. Given a distribution $p(A,B,C,D)$, we write $p(A,B,C)$ to denote the marginal distribution on the variables A,B,C. We often write $p(AB)$ instead of $p(A,B)$ for conciseness of notation. Similarly, $p(a,b,c)$ will denote the probability according to the marginal distribution $p(A,B,C)$ and we will often write it as $p(abc)$ for conciseness.

If $\mathcal{W}$ is an event, we write $p(\mathcal{W})$ to denote its probability according to p. Given a probability space p and a random variable A, when we write $A \in \mathcal{W}$ for an event $\mathcal{W}$ we only consider events in the space of values taken by the variable A.

Given a fixed value c, we denote by $\mathbb{E}_{p(b|c)}[g(a,b,c)] := \sum_b p(b \mid c) \cdot g(a,b,c)$, the expected value of the function $g(a,b,c)$ under the distribution $p(B \mid c)$. If the probability space p is clear from the context, then we will just write $\mathbb{E}_{b|c}[g(a,b,c)]$ to denote the expectation. For a Boolean function $h(a,b)$ and a probability distribution $p(A,B)$, denoting by $\mathbf{1}[h(a,b) = 0]$ the indicator function for the event $h(a,b) = 0$, we write $p(h = 0) := \mathbb{E}_{p(ab)}[\mathbf{1}[h(a,b) = 0]]$ as the probability that h is 0 under inputs drawn from p.

We write $A - M - B$ as a shorthand to say that that the random variables A, M and B form a Markov chain, or in other words, A and B are independent given M: $p(amb) = p(m) \cdot p(a \mid m) \cdot p(b \mid m)$ for every a,b,m.

To get familiar with the notation, consider the following example. Let $A \in \{0,1\}^2$ be a uniformly distributed random variable in a probability space p. Then, $p(A)$ is the uniform distribution on $\{0,1\}^2$ and if $a = (0,0)$, $p(a) = 1/4$. Let A_1 and A_2 denote the first and second bits of A, then if $B = A_1 + A_2 \bmod 2$, then when $b = 1$, $p(A \mid b)$ is the uniform distribution on $\{(0,1),(1,0)\}$. If $a = (1,0)$, and $b = 1$, then $p(a \mid b) = 1/2$, and $p(a,b) = 1/4$. If $\mathcal{E}$ is the event that $A_1 = B$, then $p(\mathcal{E}) = 1/2$. Let $q(A) = p(A \mid \mathcal{E})$, then $q(A)$ is the uniform distribution on $\{(0,0),(1,0)\}$ and $q(A_2)$ is the distribution over the sample space $\{0,1\}$ which takes the value 0 with probability 1.

2.2 Statistical distance

For two distributions $p(A), q(A)$, the statistical distance $|p(A) - q(A)|$ between them is defined to be $|p(A) - q(A)| = \max_{\mathcal{Q}} (p(A \in \mathcal{Q}) - q(A \in \mathcal{Q}))$ where $\mathcal{Q}$ is an event.

Proposition 2.1. $|p(A) - q(A)| = \dfrac{1}{2}\sum_a |p(a) - q(a)| = \sum_{a:p(a)>q(a)} (p(a) - q(a))$.

We say $p(A)$ and $q(A)$ are ε-close if $|p(A) - q(A)| \le \varepsilon$ and we write it as $p(A) \overset{\varepsilon}{\approx} q(A)$.

Proposition 2.2. *If $p(AB), q(AB)$ are such that $p(A) = q(A)$, then*

$$|p(B) - q(B)| = \mathop{\mathbb{E}}_{p(a)} [|p(B \mid a) - q(B \mid a)|] .$$

2.3 Divergence and mutual information

The *divergence* between distributions $p(A)$ and $q(A)$ is defined to be

$$\frac{p(A)}{q(A)} = \sum_a p(a) \log \frac{p(a)}{q(a)}.$$

For three random variables A, B, C and an event $\mathcal{E}$ in a probability space p, we will use the shorthand

$$\frac{A \mid bc\mathcal{E}}{A \mid c} = \frac{p(A \mid bc\mathcal{E})}{p(A \mid c)},$$

when p is clear from context. The *mutual information* between A, B conditioned on C is defined as

$$\mathbf{I}(A : B \mid C) = \mathop{\mathbb{E}}_{c,b}\left[\frac{A \mid bc}{A \mid c}\right] = \mathop{\mathbb{E}}_{c,a}\left[\frac{B \mid ac}{B \mid c}\right] = \sum_{a,b,c} p(abc) \log \frac{p(a \mid bc)}{p(a \mid c)}.$$

We shall often work with multiple probability spaces over the same sample space. To avoid confusion, we shall explicitly write $\mathbf{I}_p(A : B \mid C)$ to specify the probability space p being used for computing the mutual information.

2.4 Basic divergence facts

The proofs of the following basic facts can be found in the book by Cover and Thomas [10]. In the following, p and q are probability spaces (over the same sample space), and A, B and C are random variables on the underlying sample space.

Proposition 2.3. *If $A \in \{0,1\}^\ell$, then $\mathbf{I}(A : B) \leq \ell$.*

Proposition 2.4. *If B determines C, then $\mathbf{I}(A : C) \leq \mathbf{I}(A : B)$.*

Proposition 2.5 (Chain Rule)**.** *If $A = (A_1, \ldots, A_s)$, then* $\dfrac{p(A)}{q(A)} = \displaystyle\sum_{i=1}^{s} \mathop{\mathbb{E}}_{p(a)}\left[\frac{p(A_i \mid a_{<i})}{q(A_i \mid a_{<i})}\right].$

Proposition 2.6 (Pinsker's Inequality)**.** $|p(A) - q(A)|^2 \leq \dfrac{\ln 2}{2} \cdot \dfrac{p(A)}{q(A)} \leq \dfrac{p(A)}{q(A)}.$

Proposition 2.7. $\dfrac{p(A)}{q(A)} \geq 0.$

2.5 Divergence inequalities

The following propositions bound the change in divergence when extra conditioning is involved. Some of these were proved in [6, 15]. For completeness, we include full proofs for all of them. Below, p and q are probability spaces, and A, M and B are random variables on the underlying sample space.

Proposition 2.8. *If A and B are independent and $A - M - B$, then $\mathbf{I}(A : M) = \mathbf{I}(A : M \mid B)$.*

Proof. Since A and B are independent $p(A \mid b) = p(A)$ and since $A - M - B$, $p(A \mid mb) = p(A \mid m)$. So, we have

$$\mathbf{I}(A : M \mid B) = \mathop{\mathbb{E}}_{bm}\left[\frac{A \mid mb}{A \mid b}\right] = \mathop{\mathbb{E}}_{bm}\left[\frac{A \mid m}{A}\right] = \mathbf{I}(A : M)\,. \qquad \square$$

Proposition 2.9 ([6])**.** *For an event $\mathcal{W}$ and variables A, B in a probability space p, we have*

$$\mathop{\mathbb{E}}_{b \mid \mathcal{W}}\left[\frac{A \mid b\mathcal{W}}{A}\right] \le \log\frac{1}{p(\mathcal{W})} + \mathbf{I}(A : B \mid \mathcal{W})\,.$$

Proof. We can write

$$\begin{aligned}
\mathop{\mathbb{E}}_{b \mid \mathcal{W}}\left[\frac{A \mid b\mathcal{W}}{A}\right] - \mathbf{I}(A : B \mid \mathcal{W}) &= \sum_{ab} p(ab \mid \mathcal{W}) \log\frac{p(a \mid b\mathcal{W}) \cdot p(a \mid \mathcal{W})}{p(a) \cdot p(a \mid b\mathcal{W})} \\
&= \sum_{a} p(a \mid \mathcal{W}) \log\frac{p(a \mid \mathcal{W})}{p(a)} \\
&= \sum_{a} p(a \mid \mathcal{W}) \log\frac{p(\mathcal{W} \mid a)}{p(\mathcal{W})} \le \log\frac{1}{p(\mathcal{W})}\,. \qquad \square
\end{aligned}$$

Proposition 2.10 ([6])**.** $\mathbb{E}_{p(b)}\left[\frac{p(A \mid b)}{q(A)}\right] \ge \frac{p(A)}{q(A)}$.

Proof.

$$\mathop{\mathbb{E}}_{p(b)}\left[\frac{p(A \mid b)}{q(A)}\right] - \frac{p(A)}{q(A)} = \sum_{a,b} p(ab) \log\frac{p(a \mid b) \cdot q(a)}{q(a) \cdot p(a)} = \mathop{\mathbb{E}}_{p(b)}\left[\frac{p(A \mid b)}{p(A)}\right] \ge 0\,,$$

where the last inequality follows from Proposition 2.7. $\square$

Proposition 2.11 ([15])**.** $\mathbb{E}_{p(b)}\left[\frac{p(A \mid b)}{p(A)}\right] \le \mathbb{E}_{p(b)}\left[\frac{p(A \mid b)}{q(A)}\right]$.

Proof.

$$\mathop{\mathbb{E}}_{p(b)}\left[\frac{p(A \mid b)}{q(A)}\right] - \mathop{\mathbb{E}}_{p(b)}\left[\frac{p(A \mid b)}{p(A)}\right] = \sum_{a,b} p(ab) \log\frac{p(a \mid b) \cdot p(a)}{q(a) \cdot p(a \mid b)} = \frac{p(A)}{q(A)} \ge 0\,,$$

where the last inequality follows from Proposition 2.7. $\square$

Proposition 2.12. *Let $A \in \{0,1\}$ and let $\gamma = p(a)$ and $\varepsilon = q(a)$ for $a = 1$. Then,*

$$\frac{p(A)}{q(A)} \ge \gamma \log\frac{\gamma}{\mathrm{e}\varepsilon}\,.$$

Proof. We have

$$\frac{p(A)}{q(A)} = \gamma \log \frac{\gamma}{\varepsilon} + (1-\gamma)\log\frac{1-\gamma}{1-\varepsilon} \geq \gamma \log \frac{\gamma}{\varepsilon} + (1-\gamma)\log(1-\gamma) \geq \gamma \log \frac{\gamma}{\varepsilon} - \gamma \log \mathrm{e} = \gamma \log \frac{\gamma}{\mathrm{e}\varepsilon},$$

where in the last inequality we used the fact that

$$-(1-\gamma)\ln(1-\gamma) = (1-\gamma)\ln\left(1+\frac{\gamma}{1-\gamma}\right) \leq (1-\gamma)\frac{\gamma}{1-\gamma} = \gamma. \qquad \square$$

2.6 Communication Complexity and Information Cost

The *Communication Complexity* of a protocol is the maximum number of bits that may be exchanged by the protocol. Under an input distribution $p(X,Y)$, the *Internal Information Cost* of a randomized protocol (with both shared and private randomness) is defined to be $\mathbf{I}_p(X : M \mid YR) + \mathbf{I}_p(Y : M \mid XR)$ where X and Y are inputs to the protocol, M denotes the messages of the protocol and R denotes the shared randomness of the protocol.

We briefly describe basic properties of communication protocols that we need. For more details see the book by Kushilevitz and Nisan [20]. For a deterministic protocol π, let $\pi(x,y)$ denote the messages of the protocol on inputs x,y. For any transcript m of the protocol, define the following events:

$$\mathcal{S}_m = \{x \mid \exists y \text{ such that } \pi(x,y) = m\}, \qquad \mathcal{T}_m = \{y \mid \exists x \text{ such that } \pi(x,y) = m\}.$$

We then have:

Proposition 2.13 (Messages correspond to rectangles)**.** *If m is a transcript and x,y are inputs to a deterministic protocol π, then,*

$$\pi(x,y) = m \iff x \in \mathcal{S}_m \wedge y \in \mathcal{T}_m.$$

Proposition 2.13 implies:

Proposition 2.14 (Markov property of protocols)**.** *Let X and Y be random inputs to a deterministic protocol and let M denote the messages of this protocol. If X and Y are independent then $X - M - Y$.*

Note that the above implies that if x and y are independent inputs sampled from a distribution $p(X,Y)$ and m is a transcript of a deterministic communication protocol, then

$$p(xy \mid m) = p(xy \mid x \in \mathcal{S}_m, y \in \mathcal{T}_m) = p(x \mid x \in \mathcal{S}_m)p(y \mid y \in \mathcal{T}_m).$$

Lemma 2.15 (Errors and statistical distance)**.** *Let $h(x,y)$ be a Boolean function and $p(X,Y)$ be a distribution such that $p(h=0) = p(h=1) = 1/2$. If π is a deterministic protocol with messages M that computes h with error δ on the distribution p, then $|p(M \mid h=0) - p(M \mid h=1)| \geq 1-2\delta$.*

Proof. Since $|p(M \mid h=0) - p(M \mid h=1)| = \max_{\mathcal{Q}}(p(M \in \mathcal{Q} \mid h=0) - p(M \in \mathcal{Q} \mid h=1))$ it suffices to exhibit an event $\mathcal{Q}$ such that $p(M \in \mathcal{Q} \mid h=0) - p(M \in \mathcal{Q} \mid h=1) = 1-2\delta$. Let $\mathcal{M}_0$ denote the event that the protocol outputs a zero. Then, since $p(h=0)=p(h=1)=1/2$, writing the probability of success in terms of $\mathcal{M}_0$, we have

$$1-\delta = \frac{p(M\in\mathcal{M}_0 \mid h=0)}{2} + \frac{1-p(M\in\mathcal{M}_0 \mid h=1)}{2} = \frac{1}{2} + \frac{p(M\in\mathcal{M}_0 \mid h=0)-p(M\in\mathcal{M}_0 \mid h=1)}{2}.$$

On rearranging, the above gives us that $p(M\in\mathcal{M}_0 \mid h=0)-p(M\in\mathcal{M}_0 \mid h=1)=1-2\delta$ and hence the statistical distance must be at least $1-2\delta$. $\square$

3 The k-ary pointer jumping function

For a parameter $k\in\mathbb{N}, k\geq 2$, we work with the alphabet $[k]=\{1,2,\dots,k\}$. Let $X,Y:[k]^{<n}\to[k]$ be functions mapping strings of length less than n to a single character. Let $F,G:[k]^n\to\{0,1\}$ be Boolean functions. For $z\in[k]^n$, let $z_{\leq r}$ denote the prefix of z of length r. In the k-ary pointer jumping problem, the first party is given X,F, and the second is given Y,G. The goal of the parties is to compute $F(z)+G(z)$ mod 2, where $z\in[k]^n$ is the unique string satisfying the n equations

$$X(z_{\leq r})+Y(z_{\leq r}) = z_{r+1} \bmod k\,,$$

for every $r\in\{0,1,\dots,n-1\}$.

3.1 Input and fooling distributions

For $z\in[k]^{<n}, J\in\{0,1,\dots,n-1\}$ and X,Y as above, we say z is *consistent* with X,Y,J, if $|z|\geq J+1$, and

$$X(z_{\leq J})+Y(z_{\leq J}) = z_{J+1} \bmod k\,.$$

The distribution on inputs is described in Figure 1.

Fooling Distribution $p(X,F,Y,G)$**:** Index $J\in\{0,\dots,n-1\}$ is sampled uniformly at random. Functions $X,Y:[k]^{<n}\to[k]$ are sampled uniformly at random subject to the constraint that for any $z\in[k]^{<J}$, $X(z)=Y(z)$. Functions $F,G:[k]^n\to\{0,1\}$ are uniformly random.

Input Distribution $q(X,F,Y,G)$**:** Let $\mathcal{E}_0$ denote the event that for all consistent z, $X(z)=Y(z)$ and $F(z)=G(z)$ (when $|z|=n$). Let $\mathcal{E}_1$ denote the event that for all consistent z, $X(z)=Y(z)$ and $F(z)\neq G(z)$ (when $|z|=n$). In the distribution $q_0(X,F,Y,G)$, J is sampled uniformly from $\{0,1,\dots,n-1\}$, and the rest of the variables are sampled according to the distribution of $p(X,F,Y,G \mid \mathcal{E}_0,j)$. In the distribution $q_1(X,F,Y,G)$, J is sampled uniformly, and the rest of the variables are sampled uniformly from the distribution $p(X,F,Y,G \mid \mathcal{E}_1,j)$. The input is sampled by sampling from q_0 with probability $1/2$ and from q_1 with probability $1/2$.

Figure 1: Distributions for the k-ary pointer jumping problem.

The distribution on inputs, described in Figure 1, ensures that $F(z)+G(z)$ mod 2 is the same for *every* consistent z, so it is enough for the parties to find a consistent z to complete the goal.

Comparison with the bursting noise function. We remark that although our formulation allows us to define the k-ary pointer jumping function and the input distribution much more compactly than the bursting noise function defined in [15], our function is just a simpler variant of their construction. The main difference being that our function can be thought of as being computed over a k-ary tree as opposed to a binary tree (as in the work of [15]) and it is symmetric with respect to both parties since we are taking the sum modulo k of the values computed by both parties (as opposed to each party going down the tree alternately as in the work of [15]). Our distribution is also very similar to the distribution used by [15].

3.2 Simple protocols

To get a better understanding of the above function, let us present a few simple protocols for it.

Trivial protocol. There is a trivial protocol for this problem that has worst-case communication $O(n \log k)$: in each step Alice and Bob send each other the values $X(z_{\leq r}), Y(z_{\leq r})$, until they have computed z. The parties can compute $F(z)+G(z) \bmod 2$ with two more bits of communication. Note that this protocol succeeds with zero error under any input distribution.

Binary search protocol. Under the input distribution $q(X,F,Y,G)$, for any $z \in [k]^n$, we have that $X(z_{<J}) = Y(z_{<J})$ with probability 1 while $X(z_{\leq J})$ and $Y(z_{\leq J})$ are different with probability $1-1/k$. The players can use a version of binary search [11] to find the first difference and therefore, with probability at least $1-(\varepsilon+1/k)$, they can compute the index J with $O(\log((n\log k)/\varepsilon))$ bits of communication. With an additional $2\log k$ bits of communication the players can then find a consistent z (satisfying $X(z_{\leq J}) + Y(z_{\leq J}) = z_{J+1} \bmod k$) which suffices to compute the value of the function on the input distribution $q(X,F,Y,G)$. This protocol has communication $O(\log((n\log k)/\varepsilon)+\log k)$ and the error probability is at most $\varepsilon+1/k$ on the input distribution $q(X,F,Y,G)$.

Sampling protocol. Let $t = \Theta\left(k^2 \log(1/\varepsilon)\right)$. Using shared randomness, the players draw a subset $\mathcal{S}$ by choosing t strings uniformly at random from $[k]^n$, exchange the values of $F(z)$ and $G(z)$ for each $z \in \mathcal{S}$ and output the majority of the bits $\{F(z)+G(z) \bmod 2\}_{z\in\mathcal{S}}$. Under the input distribution $q(X,F,Y,G)$, the probability that a random string is consistent (that it satisfies $X(z_{\leq J}) + Y(z_{\leq J}) = z_{J+1} \bmod k$) is $1/k$, so using standard concentration bounds, this protocol has error at most ε under the distribution $q(X,F,Y,G)$. Moreover, the communication is $O(k^2\log(1/\varepsilon))$.

3.3 Information and Communication Complexity

We prove that there is a low-information solution for this task, with internal information cost $O\left((\log k + \log\log n)\cdot 2^{(2\log n)/k}\right)$ on the distribution $q(X,F,Y,G)$ (Theorem 1.1) but any randomized protocol that errs with a constant probability on the input distribution $q(X,F,Y,G)$ requires communication at least $\Omega(\min\{k, \log n\})$ (Theorem 1.2). The lower bound is tight up to polynomial factors, as the binary search and sampling protocol described previously, show that the communication complexity is $O(\min\{k^2, \log n + \log k\})$. Setting $n = 2^k$, we get a correct protocol with information cost $O(\log k)$ even though no protocol with communication $\Omega(k)$ can succeed.

Low-information protocol. The low-information protocol for the problem is quite similar to the trivial protocol, so let us first discuss the information cost of the trivial protocol under the distribution $q(X,F,Y,G)$. At the beginning of the protocol, with just their own inputs in hand, the parties do not have any information about J. Using the trivial protocol, both parties learn the value of J with high probability. This happens because J is close to the first point at which their inputs disagree. Since the entropy of J is $\Theta(\log n)$ bits and J is determined with high probability given X and Y, it is not too hard to argue that the parties learn $\Omega(\log n)$ bits of information about each other's inputs. It follows that the internal information cost of the trivial protocol is $\Omega(\log n)$ bits, much larger than what we are aiming for.

The low-information protocol adds some noise to hide the value of J. In each step of the low-information protocol, the parties send each other the value $X(z_{\leq r}), Y(z_{\leq r})$ with probability $1-(1/\log n)$ and send a uniformly random value otherwise. The parties abort the protocol if they experience $\log n/\log\log n$ rounds where the messages they sent were not the same. The distribution on inputs ensures that they will sample a consistent z with high probability. When the parties sample a consistent z, the messages sent are almost always sampled from a distribution that the receiving party knows, while if they sample a z that is not consistent, the protocol aborts shortly after the inconsistency. These properties can be used to show that under the distribution $q(X,F,Y,G)$, the information cost of the protocol is $O((\log k+\log\log n)\cdot 2^{(2\log n)/k})$ and the error probability is at most $4/\log n$.

Intuitively, this protocol does not reveal a lot of information about J because at the end of the protocol the parties see a lot of disagreements, and so they only learn that J belongs to some set of density $1/\log n$. Heuristically, the entropy of J conditioned on the entire exchange is $\approx \log(n/\log n) = \log n - \log\log n$ and the amount of information revealed about J is $O(\log\log n)$.

In Section 5, we analyze the aforementioned low-information protocol and prove Theorem 1.1. Before moving on, we remark that when $n=2^k$, using this low information-cost protocol with the simulation result of Braverman [3], one can obtain a deterministic protocol with communication complexity k^{c/ε^2} for a constant $c>2$, that computes the k-ary pointer jumping function with error probability $O(\varepsilon+1/k)$ on the distribution $q(X,F,Y,G)$.

Communication lower bound. Consider any protocol with ℓ bits of communication that solves the k-ary pointer jumping problem. Without loss of generality, we may assume that the protocol is deterministic, since any randomness can always be fixed to obtain a deterministic protocol that succeeds with high probability. Let M denote the messages of the protocol. Our input distribution $q(X,F,Y,G)$ has the property that under the inputs drawn from this distribution the k-ary pointer jumping function h is balanced (it takes values 1 and 0 with probability half each). Hence, if the protocol solved the k-ary pointer jumping problem with error δ on the distribution $q(X,F,Y,G)$, then the statistical distance between $q_0(M)$ (the induced distribution on M when inputs are drawn from $q(X,F,Y,G)$ conditioned on the value of k-ary pointer jumping function being 0) and $q_1(M)$ (the induced distribution on M when inputs are drawn from $q(X,F,Y,G)$ conditioned on the value of k-ary pointer jumping function being 1) would be at least $1-2\delta$ (see Lemma 2.15), since these distributions have nearly disjoint supports.

To prove the communication lower bound, we define the *fooling distribution* $p(X,F,Y,G)$ on inputs, and using information theoretic techniques show that if the communication complexity ℓ of the protocol is much less than $\min\{k,\log n\}$, then both of the distributions $q_0(M)$ and $q_1(M)$ are close to the fooling distribution $p(M)$, which implies that the statistical distance between $q_0(M)$ and $q_1(M)$ is close to 0. This will give us a contradiction, since we argued in the paragraph above that these distributions must be far apart.

It will be convenient to state our results in terms of the function $\eta : [0,\infty) \to [0,1]$ defined as

$$\eta(\alpha) = \begin{cases} 0 & \text{if } \alpha = 0, \\ \alpha \log(1/\alpha) & \text{if } \alpha \in (0, 1/\mathrm{e}), \\ \log(\mathrm{e})/\mathrm{e} & \text{if } \alpha \geq 1/\mathrm{e}. \end{cases} \tag{3.1}$$

One can check that η is non-decreasing, continuous, and concave. We prove the following upper bounds on statistical distance between the induced message distributions.

Theorem 3.1. *For any deterministic protocol for the k-ary pointer jumping function with communication complexity at most ℓ, we have that*

$$q_0(M) \overset{\gamma}{\approx} p(M) \overset{\gamma}{\approx} q_1(M), \text{ with } \gamma = 4\sqrt[3]{\left(2\mathrm{e}\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right)\right)}.$$

We stress that the fooling distribution $p(X,F,Y,G)$ is only used in the analysis. The inputs to the protocol come from the distribution $q(X,F,Y,G)$.

Theorem 3.1 implies that $\ell = \Omega(\min\{k, \log n\})$ for any protocol that solves the k-ary pointer jumping problem for the input distribution $q(X,F,Y,G)$. This gives us Theorem 1.2.

Proof of Theorem 1.2. Consider any protocol for the k-ary pointer jumping function that has communication complexity $\ell := \varepsilon^3 \cdot \min\{k, \log n\}$ and errs with probability δ on the input distribution $q(X,F,Y,G)$. We may assume without loss of generality that the protocol is deterministic, since if the protocol used public or private randomness, we can fix the randomness to get a deterministic protocol with the same communication complexity and the same error on the input distribution $q(X,F,Y,G)$. Since, the protocol has error at most δ, the statistical distance between $q_0(M)$ and $q_1(M)$ must be at least $1-2\delta$ (see Lemma 2.15). On the other hand, Theorem 3.1 implies that

$$|q_0(M) - q_1(M)| \leq |q_0(M) - p(M)| + |p(M) - q_1(M)| \leq 2\gamma,$$

where

$$\gamma = 4\sqrt[3]{\left(2\mathrm{e}\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right)\right)} < 8\varepsilon$$

for large enough values of k and n. Then, it must be that $1 - 2\delta \leq 2\gamma < 16\varepsilon$ and hence, the error $\delta > 1/2 - 8\varepsilon$. □

4 Communication lower bound

4.1 High-level proof sketch for Theorem 3.1

Consider an ℓ-bit deterministic protocol. Our goal is to show that if $\ell \ll \min\{k, \log n\}$, then the induced distribution of the messages M of the protocol is roughly the same under the fooling distribution

$p(X,F,Y,G)$ and the distribution $q_0(X,F,Y,G)$ (input distribution $q(X,F,Y,G)$ conditioned on the value of k-ary pointer jumping function being 0) and that a similar statement is true for $p(X,F,Y,G)$ and $q_1(X,F,Y,G)$. As described in the previous section, this proves that communication complexity must be $\Omega(\min\{k,\log n\})$.

How are the distributions $p(X,F,Y,G)$ and $q_0(X,F,Y,G)$ related? Let S denote the set of consistent strings $z \in [k]^{\leq n}$ and $X_S Y_S F_S G_S$ denote the projection of the random variables $XYFG$ on the strings in the set S and let $X_{\leq J}, Y_{\leq J}$ denote the restriction of X,Y to inputs of length at most J. Let X_J, Y_J denote the restriction of X,Y to inputs of length J.

First note that the set of consistent strings S is determined by $X_J Y_J J$, since given $X_J Y_J J$, one can check whether any string $z \in [k]^n$ is consistent or not (whether it satisfies $X(z_{\leq J}) + Y(z_{\leq J}) = z_{J+1} \mod k$). Since the distribution $p(X,F,Y,G)$ and $q_0(X,F,Y,G)$ are the same on $X_{\leq J} Y_{\leq J} J$, it follows that the distributions $p(S)$ and $q_0(S)$ are also the same. Fixing the values $X_{\leq J} Y_{\leq J} J$, the distribution $q_0(X,F,Y,G)$ is obtained from $p(X,F,Y,G)$ by conditioning on the event $X_S F_S = Y_S G_S$. Similarly after fixing $X_{\leq J} Y_{\leq J} J$, the distribution $q_1(X,F,Y,G)$ is obtained from $p(X,F,Y,G)$ by conditioning on the event $X_S F_S = Y_S \overline{G}_S$, where $\overline{G}$ is the function $1-G$.

So, after fixing $X_{\leq J} Y_{\leq J} J$, to prove that $p(M) \approx q_0(M)$, we need to show that the distribution $p(M)$ does not change by much, even if we condition on the event $X_S F_S = Y_S G_S$ (and similarly $X_S F_S = Y_S \overline{G}_S$). We prove this in three steps:

- First, we argue using a subtle application of the chain rule, that if $\ell \ll \min\{k,\log n\}$, the protocol does not reveal much information about S under the *fooling distribution* $p(X,F,Y,G)$.
- Next, we show that since the players do not learn much information about S, they also do not learn much information about $X_S F_S$ and $Y_S G_S$ in the *fooling distribution* $p(X,F,Y,G)$. To argue this, we prove a generalization of Shearer's Lemma [9, 23].
- Lastly, using the above facts we show that the distribution $p(M)$ roughly stays the same even after conditioning on the event $X_S F_S$ and $Y_S G_S$.

Note that it is only during the last step that the input distribution $q_0(X,F,Y,G)$ (and $q_1(X,F,Y,G)$) are related to the fooling distribution $p(X,F,Y,G)$. The first two steps analyze the behavior of the protocol only on the fooling distribution $p(X,F,Y,G)$. Next we elaborate on each of the steps.

Information revealed about consistent strings

As discussed before, the set S of consistent strings is determined by $X_J Y_J J$ in the fooling distribution $p(X,F,Y,G)$. We bound the amount of information revealed about S to each party as follows.

Lemma 4.1. *Let M denote the messages of a deterministic ℓ-bit protocol, then*

$$\left.\begin{aligned} \mathbf{I}_p(M:S \mid Y_{\leq J}J) \leq \mathbf{I}_p(M:X_J \mid Y_{<J}J) \\ \mathbf{I}_p(M:S \mid X_{\leq J}J) \leq \mathbf{I}_p(M:Y_J \mid X_{<J}J) \end{aligned}\right\} \leq \frac{2^\ell}{n}.$$

Since in the distribution $p(X,F,Y,G)$, we have $X_{<J} = Y_{<J}$, it follows that $\mathbf{I}_p(M:X_J \mid Y_{<J}J) = \mathbf{I}_p(M:X_J \mid X_{<J}J)$. At first glance, one might think that the expression $\mathbf{I}_p(M:X_J \mid X_{<J}J)$ can be upper bounded by ℓ/n using the chain rule (by averaging over all n values of J). However, this is not true:

since J is essentially determined by X and Y and the messages M contain information about X and Y, it is not clear if one can directly use the chain rule as simply as above. To understand this point in more detail, it is worthwhile to consider a simple example.

Consider the following variant of the binary search protocol for the k-ary pointer jumping problem from Section 3.2. In that section, we discussed the protocol on the input distribution $q(X,F,Y,G)$, however the binary search phase of the protocol to find the index J works pretty much the same on the fooling distribution $p(X,F,Y,G)$. Using $O(\log n)$ bits the players determine the value of J with a version of binary search and then exchange one bit about the values of X_J and Y_J. So, in this case the $O(\log n)$-bit protocol reveals at least 1 bit of information about X_J and Y_J. So, one could not hope to get a bound like ℓ/n since this $O(\log n)$-bit protocol already reveals 1 bit of information. In fact, this protocol also shows that the above lemma is tight. If we stop the binary search phase after $O(\ell)$ bits, the players will know a set of size $n/2^\ell$ in which J lies. If the players choose a random index K in this set and reveal one bit of information about X_K, then the amount of information revealed about X_J is exactly $2^\ell/n$. as the players reveal one bit about X_J with probability $2^\ell/n$.

Despite the above, we are able to prove Lemma 4.1 by a subtle application of the chain rule and the Markov property of protocols. Essentially, the proof argues that for each of the 2^ℓ possible transcripts, the protocol, on average, only reveals $1/n$ bit of information.

Information revealed about X_SF_S and Y_SG_S

Next, we want show that the parties do not learn much information about the values of X,F,Y,G restricted to S (denoted X_S,F_S,Y_S,G_S). Note that only $1/k$ fraction of the strings are in S, since for every z, $p(z \in S \mid J = j, X_{\leq J} = x_{\leq j}) \leq 1/k$. So, recalling the classical Shearer's Lemma [9, 23], one might hope that $\mathbf{I}(M : X_SF_S)$ can be bounded by $\mathbf{I}(M : XF)/k \leq \ell/k$. If S was independent of M,X,F, such a bound would be easy to prove using the chain rule for information (see Lemma 8 in [15]).

In our case, S is not independent of M but almost independent, as the amount of information that M reveals about S is small (after conditioning on $X_{<J}J$). So it is conceivable that $\mathbf{I}(M : X_SF_S)$ can still be bounded by $O(\mathbf{I}(M : XF)/k)$ plus some small error terms. We prove such a generalization of Shearer's Lemma which may be of independent interest. Below U_S denotes the restriction of U to the coordinates in S. We prove the following lemma.

Lemma 4.2. *Given a probability space p', let $U = (U_1,\ldots,U_t)$ where $U_1,\ldots,U_t$ are mutually independent random variables. Let random variables $C \in \{0,1\}^\ell, S \subseteq [t]$, and V be such that U is independent of SV, and $U - C - SV$ and for all $i \in [t]$, $p'(i \in S) \leq 1/k$. Then*

$$\mathbf{I}(C : U_S \mid VS) \leq \ell \cdot \left(\frac{2\mathrm{e}}{k} + 2\sqrt{\mathbf{I}(C:S)}\right) + \eta\left(\sqrt{\mathbf{I}(C:S)}\right).$$

Conditioned on any fixing of $X_{\leq J}Y_{\leq J}J$, we have that XF and YG are independent under p, and since M denotes the messages in a communication protocol, the Markov property (Proposition 2.14) implies that $XF - M - YG$ holds. Thus, Lemma 4.2 in conjunction with Lemma 4.1 and convexity can be used to show that the amount of information the messages reveal about X_SF_S (and similarly for Y_SG_S) is small:

$$\left.\begin{matrix}\mathbf{I}_p(M : X_SF_S \mid X_{\leq J}Y_{\leq J}J) \\ \mathbf{I}_p(M : Y_SG_S \mid X_{\leq J}Y_{\leq J}J)\end{matrix}\right\} \leq 2\mathrm{e}\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right). \tag{4.1}$$

Conditioning on the event $X_SF_S = Y_SG_S$

Lastly, we need to show that the distribution $p(M)$ does not change much after conditioning on the event $X_SF_S = Y_SG_S$ which gives us the distribution $q_0(M)$. Since, both events are essentially independent of M (the mutual information as in (4.1) is small), one can expect that conditioning on the event $X_SF_S = Y_SG_S$ should not change the distribution of M by much.

In fact, we show the following general statement.

Lemma 4.3. *Given a probability space p', if $A,B \in [T]$ are uniform and independent random variables, and $A-C-B$, then*

$$p'(C) \stackrel{\varepsilon}{\approx} p'(C \mid A=B), \quad \textit{with} \quad \varepsilon = 2\sqrt[3]{\mathbf{I}(C:A)} + 2\sqrt[3]{\mathbf{I}(C:B)}.$$

Lemma 4.3 together with (4.1) and another convexity argument completes the proof of Theorem 3.1.

4.2 Proof of Theorem 3.1

We shall prove that $p(M) \stackrel{\gamma}{\approx} q_0(M)$. The bound for the distribution $q_1(M)$ is analogous. We first give the proof assuming Lemmas 4.1, 4.2, and 4.3 and then we prove the lemmas.

By Lemma 4.1, $\mathbf{I}_p(M:S \mid X_{\leq J}J) \leq 2^\ell/n$. After fixing $x_{\leq j}, j$, S is determined by Y_J. For any such fixing, we have that $p(z \in S \mid x_{\leq j}j) \leq 1/k$ holds for any string z, and that XF is independent of YG. Furthermore, by Proposition 2.14 we also have $XF - M - SY_J$ after fixing $x_{\leq j}, j$. Thus we can apply Lemma 4.2 with $U = XF$ and $C = M$, $V = Y_J$, to conclude that

$$\mathbf{I}_p(M:X_SF_S \mid SY_Jx_{\leq j}j) \leq 2\mathrm{e}\ell/k + 2\ell\sqrt{\mathbf{I}_p(M:S \mid x_{\leq j}j)} + \eta\left(\sqrt{\mathbf{I}_p(M:S \mid x_{\leq j}j)}\right).$$

Taking the expectation over the choice of $x_{\leq j}j$, and using the concavity of the functions $\sqrt{\alpha}$ and $\eta(\alpha)$, it follows that

$$\begin{aligned}\mathbf{I}_p(M:X_SF_S \mid Y_{\leq J}X_{\leq J}J) &\leq 2\mathrm{e}\ell/k + 2\ell\sqrt{\mathbf{I}_p(M:S \mid X_{\leq J}J)} + \eta\left(\sqrt{\mathbf{I}_p(M:S \mid X_{\leq J}J)}\right)\\ &\leq 2\mathrm{e}\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right).\end{aligned}$$

The same bound applies to $\mathbf{I}_p(M:Y_SG_S \mid Y_{\leq J}X_{\leq J}J)$. For each fixing of $x_{\leq j}y_{\leq j}j$, we have $X_SF_S - M - Y_SG_S$. Thus we can apply Lemma 4.3 to conclude that

$$\begin{aligned}&|p(M \mid x_{\leq j}y_{\leq j}j) - p(M \mid x_{\leq j}y_{\leq j}j, X_SF_S = Y_SG_s)| \leq\\ &\qquad 2\sqrt[3]{\mathbf{I}_p(M:X_SF_S \mid x_{\leq j}y_{\leq j}j)} + 2\sqrt[3]{\mathbf{I}_p(M:Y_SG_S \mid x_{\leq j}y_{\leq j}j)}.\end{aligned}$$

Since $p(X_{\leq J}Y_{\leq J}J) = q_0(X_{\leq J}Y_{\leq J}J)$, we can use Proposition 2.2 to bound

$$\begin{aligned}
|p(M) - q_0(M)| &= \mathop{\mathbb{E}}_{p(x_{\leq j}y_{\leq j}j)} \left[|p(M \mid x_{\leq j}y_{\leq j}j) - p(M \mid x_{\leq j}y_{\leq j}j, X_S F_S = Y_S G_S)| \right] \\
&\leq \mathop{\mathbb{E}}_{p(x_{\leq j}y_{\leq j}j)} \left[2\sqrt[3]{\mathbf{I}_p\left(M : X_S F_S \mid x_{\leq j}y_{\leq j}j\right)} + 2\sqrt[3]{\mathbf{I}_p\left(M : Y_S G_S \mid x_{\leq j}y_{\leq j}j\right)} \right] \\
&\leq 2\sqrt[3]{\mathbf{I}_p\left(M : X_S F_S \mid X_{\leq J}Y_{\leq J}J\right)} + 2\sqrt[3]{\mathbf{I}_p\left(M : Y_S G_S \mid X_{\leq J}Y_{\leq J}J\right)} \\
&\leq 4\sqrt[3]{2\mathrm{e}\ell/k + 2\ell\sqrt{2^\ell/n} + \eta\left(\sqrt{2^\ell/n}\right)} = \gamma,
\end{aligned}$$

where the second to last inequality follows from the concavity of $\sqrt[3]{\alpha}$ over non-negative reals.

4.2.1 Proof of Lemma 4.1

Once $Y_{\leq J}J$ are fixed, S is determined by X_J, since whether a string z is consistent or not is determined given $X_J Y_J J$. Furthermore, after $Y_{<J}J$ is fixed, XF and YG are independent in the distribution $p(X,F,Y,G)$, so the Markov property of protocols (Proposition 2.14) implies that $X_J - M - Y_J$ (conditioned on $X_{<J}J$). Thus, using Propositions 2.4 and 2.8, we get that

$$\mathbf{I}_p\left(M : S \mid Y_{\leq J}J\right) \leq \mathbf{I}_p\left(M : X_J \mid Y_{\leq J}J\right) = \mathbf{I}_p\left(M : X_J \mid Y_{<J}J\right). \tag{4.2}$$

Since $X_{<J} = Y_{<J}$, we have that $\mathbf{I}_p\left(M : X_J \mid Y_{<J}J\right) = \mathbf{I}_p\left(M : X_J \mid X_{<J}J\right)$, which we shall show is at most $2^\ell/n$. Recall that any message m induces a rectangle $\mathcal{S}_m \times \mathcal{T}_m$ in the input space as given by Proposition 2.13. Denoting by $\mathcal{S}_m$ (and $\mathcal{T}_m$) the event that $X \in \mathcal{S}_m$ (and $Y \in \mathcal{T}_m$), Proposition 2.13 implies that $M = m$ is equivalent to the events $\mathcal{S}_m \wedge \mathcal{T}_m$. Also, since XF and YG are independent given $x_{<j}j$, by Proposition 2.14, we have $p(X \mid m, x_{<j}j) = p(X \mid \mathcal{S}_m \wedge \mathcal{T}_m, x_{<j}j) = p(X \mid \mathcal{S}_m, x_{<j}j)$. So, we can write

$$\mathbf{I}_p\left(M : X_J \mid X_{<J}J\right) = \mathop{\mathbb{E}}_{yxj|m} \left[\frac{X_j \mid m x_{<j} j}{X_j \mid x_{<j} j} \right] = \mathop{\mathbb{E}}_{yxj|m} \left[\frac{X_j \mid \mathcal{S}_m, x_{<j} j}{X_j \mid x_{<j} j} \right]. \tag{4.3}$$

The right hand side in (4.3) can be rewritten as

$$\sum_m p(\mathcal{S}_m) p(\mathcal{T}_m \mid \mathcal{S}_m) \mathop{\mathbb{E}}_{xj|\mathcal{S}_m,\mathcal{T}_m} \left[\frac{X_j \mid \mathcal{S}_m, x_{<j} j}{X_j \mid x_{<j} j} \right] \leq \sum_m p(\mathcal{S}_m) \mathop{\mathbb{E}}_{xj|\mathcal{S}_m} \left[\frac{X_j \mid \mathcal{S}_m, x_{<j} j}{X_j \mid x_{<j} j} \right], \tag{4.4}$$

where the inequality follows from the fact that $\mathbb{E}_a\left[h(a)\right] \geq p(\mathcal{W})\,\mathbb{E}_{a|\mathcal{W}}\left[h(a)\right]$, for any non-negative function h.

Since J is independent of X, we have that $p(XJ) = p(J)p(X)$ and also $p(XJ \mid \mathcal{S}_m) = p(J)p(X \mid \mathcal{S}_m)$. So, we can write

$$\mathop{\mathbb{E}}_{xj|\mathcal{S}_m} \left[\frac{X_j \mid \mathcal{S}_m, x_{<j} j}{X_j \mid x_{<j} j} \right] = \sum_{j=1}^n p(j) \mathop{\mathbb{E}}_{x|\mathcal{S}_m} \left[\frac{X_j \mid \mathcal{S}_m, x_{<j}}{X_j \mid x_{<j}} \right].$$

Since $p(J)$ is uniform we can use the chain rule to write the right hand side above as

$$\sum_{j=1}^{n} p(j) \mathop{\mathbb{E}}_{x|\mathcal{S}_m}\left[\frac{X_j \mid \mathcal{S}_m, x_{<j}}{X_j \mid x_{<j}}\right] = \frac{1}{n}\sum_{j=1}^{n} \mathop{\mathbb{E}}_{x|\mathcal{S}_m}\left[\frac{X_j \mid \mathcal{S}_m, x_{<j}}{X_j \mid x_{<j}}\right] = \frac{1}{n}\cdot\frac{X \mid \mathcal{S}_m}{X}.$$

Plugging the above into (4.4), we get that the left hand side in (4.3) can be bounded by

$$\frac{1}{n}\sum_{m} p(\mathcal{S}_m)\cdot\frac{X \mid \mathcal{S}_m}{X} \leq \frac{1}{n}\sum_{m} p(\mathcal{S}_m)\log\frac{1}{p(\mathcal{S}_m)} \leq \frac{2^{\ell}}{n},$$

where the first inequality follows from Proposition 2.9 (with $B = \perp$ and $\mathcal{W} = \mathcal{S}_m$) and the second from the fact that for $0 \leq \gamma \leq 1$, it holds that $\gamma \log(1/\gamma) \leq \log(\mathrm{e})/\mathrm{e} < 1$.

This proves that $\mathbf{I}_p(M : S \mid Y_{\leq J} J) \leq \mathbf{I}_p(M : X_J \mid Y_{<J} J) \leq 2^{\ell}/n$. The bound on $\mathbf{I}_p(M : S \mid X_{\leq J} J)$ follows analogously.

4.2.2 Proof of Lemma 4.2

First we prove the following upper bound for the conditional information in terms of the divergence.

Claim 4.4. $\mathbf{I}(C : U_S \mid VS) \leq \sum_{i=1}^{t} \mathop{\mathbb{E}}_{cu}\left[p'(i \in S \mid c)\cdot\frac{U_i \mid cu_{<i}}{U_i \mid u_{<i}}\right].$

Call c bad if $p'(i \in S \mid c) \geq 2\mathrm{e}/k + \sqrt{\mathbf{I}(C : S)}$ for some $i \in [t]$ and let $\mathcal{W}$ denote the event that C is bad. Note that when the complement event $\overline{\mathcal{W}}$ occurs, then $p'(i \in S \mid c) \leq 2\mathrm{e}/k + \sqrt{\mathbf{I}(C : S)}$ for all i. We next show that, if $\mathbf{I}(C : S)$ is small, C is bad only with a small probability.

Claim 4.5. $p'(\mathcal{W}) \leq \sqrt{\mathbf{I}(C : S)}$.

We can now prove Lemma 4.2, using Claims 4.4 and 4.5. First, it holds that

$$\mathbf{I}(C : U_S \mid VS) \leq p'(\mathcal{W}) \sum_{i=1}^{t} \mathop{\mathbb{E}}_{cu|\mathcal{W}}\left[\frac{U_i \mid cu_{<i}}{U_i \mid u_{<i}}\right] + \left(\frac{2\mathrm{e}}{k} + \sqrt{\mathbf{I}(C : S)}\right)\cdot\sum_{i=1}^{t} \mathop{\mathbb{E}}_{cu}\left[\frac{U_i \mid cu_{<i}}{U_i \mid u_{<i}}\right]. \tag{4.5}$$

When c is bad, we have that $p'(U_i \mid cu_{<i}\mathcal{W}) = p'(U_i \mid cu_{<i})$ and so,

$$\frac{U_i \mid cu_{<i}}{U_i \mid u_{<i}} = \frac{U_i \mid cu_{<i}\mathcal{W}}{U_i \mid u_{<i}}.$$

Plugging this into the right hand side in (4.5) and using the chain rule, we get that

$$\begin{aligned}\mathbf{I}(C : U_S \mid VS) &\leq p'(\mathcal{W}) \sum_{i=1}^{t} \mathop{\mathbb{E}}_{cu|\mathcal{W}}\left[\frac{U_i \mid cu_{<i}\mathcal{W}}{U_i \mid u_{<i}}\right] + \left(\frac{2\mathrm{e}}{k} + \sqrt{\mathbf{I}(C : S)}\right)\sum_{i=1}^{t} \mathop{\mathbb{E}}_{cu}\left[\frac{U_i \mid cu_{<i}}{U_i \mid u_{<i}}\right] \\ &= p'(\mathcal{W}) \mathop{\mathbb{E}}_{c|\mathcal{W}}\left[\frac{U \mid c\mathcal{W}}{U}\right] + \left(\frac{2\mathrm{e}}{k} + \sqrt{\mathbf{I}(C : S)}\right)\mathop{\mathbb{E}}_{c}\left[\frac{U \mid c}{U}\right].\end{aligned}$$

Using Proposition 2.9 and that

$$\mathop{\mathbb{E}}_{c}\left[\frac{U \mid c}{U}\right] = \mathbf{I}(U:C)\,,$$

we can bound the above as

$$\begin{aligned}\mathbf{I}(C:U_S \mid VS) &\le p'(\mathcal{W})\left(\log\frac{1}{p'(\mathcal{W})} + \mathbf{I}(U:C\mid\mathcal{W})\right) + \left(\frac{2\mathrm{e}}{k} + \sqrt{\mathbf{I}(C:S)}\right)\cdot\mathbf{I}(U:C)\\ &\le \eta\left(p'(\mathcal{W})\right) + p'(\mathcal{W})\cdot\mathbf{I}(U:C\mid\mathcal{W}) + \sqrt{\mathbf{I}(C:S)}\cdot\mathbf{I}(U:C) + \frac{2\mathrm{e}\mathbf{I}(U:C)}{k}\,.\end{aligned} \tag{4.6}$$

Using Claim 4.5, $p'(\mathcal{W}) \le \sqrt{\mathbf{I}(C:S)}$. Since η (see (3.1)) is a non-decreasing function, we can then bound $\eta(p'(\mathcal{W})) \le \eta(\sqrt{\mathbf{I}(C:S)})$. By Proposition 2.3, $\mathbf{I}(U:C\mid\mathcal{W}), \mathbf{I}(U:C) \le \ell$, so plugging it into (4.6), we get that

$$\mathbf{I}(C:U_S\mid VS) \le \eta\left(\sqrt{\mathbf{I}(C:S)}\right) + 2\ell\sqrt{\mathbf{I}(C:S)} + \frac{2\mathrm{e}\ell}{k}\,.$$

This finishes the proof of Lemma 4.2. It only remains to prove the claims.

Proof of Claim 4.4. For $i \in [t]$, set $U_i^* = U_i$ if $i \in S$ and set $U_i^* = \perp$ otherwise. Since we have that $\mathbf{I}(C:U_S\mid VS) = \mathbf{I}(C:U_1^*U_2^*\ldots U_t^*\mid VS)$, by the chain rule, we get

$$\mathbf{I}(C:U_S\mid VS) = \mathop{\mathbb{E}}_{cvs}\left[\frac{U_1^*\ldots U_t^*\mid cvs}{U^*\mid vs}\right] = \mathop{\mathbb{E}}_{cuvs}\left[\sum_{i=1}^{t}\frac{U_i^*\mid cu_{<i}^*vs}{U_i^*\mid u_{<i}^*vs}\right] = \mathop{\mathbb{E}}_{cuvs}\left[\sum_{i\in s}\frac{U_i\mid cu_{<i}^*vs}{U_i\mid u_{<i}^*vs}\right],$$

where the last equality holds because when $i \notin S$, $U_i^* = \perp$ (and so the divergence is 0) and when $i \in S$, $U_i^* = U_i$.

By assumption, U is independent of VS, $U - C - VS$ and $p'(u_i \mid u_{<i}^*) = p'(u_i) = p'(u_i \mid u_{<i})$, so the right hand side above can be written as

$$\mathop{\mathbb{E}}_{cuvs}\left[\sum_{i\in s}\frac{U_i\mid cu_{<i}^*vs}{U_i\mid u_{<i}^*vs}\right] = \mathop{\mathbb{E}}_{cvs}\left[\mathop{\mathbb{E}}_{u|c}\left[\sum_{i\in s}\frac{U_i\mid cu_{<i}^*}{U_i\mid u_{<i}^*}\right]\right] = \mathop{\mathbb{E}}_{cvs}\left[\mathop{\mathbb{E}}_{u|c}\left[\sum_{i\in s}\frac{U_i\mid cu_{<i}^*}{U_i\mid u_{<i}}\right]\right]. \tag{4.7}$$

Let $\mathbf{1}[i\in S]$ denote the indicator variable for the event that $i \in S$. Using linearity of expectation and Proposition 2.10, we have that

$$(4.7) = \mathop{\mathbb{E}}_{cvs}\left[\sum_{i\in s}\mathop{\mathbb{E}}_{u|c}\left[\frac{U_i\mid cu_{<i}^*}{U_i\mid u_{<i}}\right]\right] \le \mathop{\mathbb{E}}_{cvs}\left[\sum_{i\in s}\mathop{\mathbb{E}}_{u|c}\left[\frac{U_i\mid cu_{<i}}{U_i\mid u_{<i}}\right]\right] = \mathop{\mathbb{E}}_{cuvs}\left[\sum_{i}\mathbf{1}[i\in s]\frac{U_i\mid cu_{<i}}{U_i\mid u_{<i}}\right],$$

where the first inequality above follows from Proposition 2.10 and the definition of U_i^* (which depends only on U_i and S).

Finally, using $U - C - VS$ once more, we get that the right hand side above equals

$$\mathop{\mathbb{E}}_{cu}\left[\sum_{i}\mathop{\mathbb{E}}_{s|c}[\mathbf{1}[i\in s]]\frac{U_i\mid cu_{<i}}{U_i\mid u_{<i}}\right] = \sum_{i=1}^{t}\mathop{\mathbb{E}}_{cu}\left[p'(i\in S\mid c)\cdot\frac{U_i\mid cu_{<i}}{U_i\mid u_{<i}}\right].\qquad\square$$

Proof of Claim 4.5. Define $S_i = 1$ if $i \in S$ and 0 otherwise. We are going to show that whenever c is bad, we have that

$$\frac{S \mid c}{S} \geq \sqrt{\mathbf{I}(C:S)}\,.$$

Since

$$\mathbf{I}(C:S) = \mathop{\mathbb{E}}_{c}\left[\frac{S \mid c}{S}\right],$$

Markov's inequality will imply that the probability of c being bad is at most $\sqrt{\mathbf{I}(C:S)}$ and hence $p(\mathcal{W}) \leq \sqrt{\mathbf{I}(C:S)}$.

When c is bad, there is an i^* such that $p'(i^* \in S \mid c) \geq 2\mathrm{e}/k + \sqrt{\mathbf{I}(C:S)}$. By chain rule and Proposition 2.7,

$$\frac{S \mid c}{S} \geq \frac{S_{i^*} \mid c}{S_{i^*}}.$$

Since $p'(i \in S) \leq 1/k$ for any i, by Proposition 2.12,

$$\begin{aligned}\frac{S \mid c}{S} \geq \frac{S_{i^*} \mid c}{S_{i^*}} &\geq p'(i^* \in S \mid c)\log\left(\frac{k \cdot p'(i^* \in S \mid c)}{\mathrm{e}}\right) \\ &\geq \left(\frac{2\mathrm{e}}{k} + \sqrt{\mathbf{I}(C:S)}\right)\log\left(\frac{k}{\mathrm{e}}\left(\frac{2\mathrm{e}}{k} + \sqrt{\mathbf{I}(C:S)}\right)\right) \geq \sqrt{\mathbf{I}(C:S)}\,.\end{aligned}$$

This finishes the proof of the claim. □

4.2.3 Proof of Lemma 4.3

We assume $\mathbf{I}(C:A), \mathbf{I}(C:B) \leq 1$, since otherwise the lemma is trivially true. For brevity, set

$$\alpha^3 = \mathbf{I}(C:A) = \mathop{\mathbb{E}}_{c}\left[\frac{A \mid c}{A}\right] \text{ and } \beta^3 = \mathbf{I}(C:B) = \mathop{\mathbb{E}}_{c}\left[\frac{B \mid c}{B}\right].$$

Call c *bad* if

$$\frac{A \mid c}{A} \geq \alpha^2 \quad \text{or} \quad \frac{B \mid c}{B} \geq \beta^2\,,$$

and good otherwise. By Markov's inequality, the probability that C is bad is at most $\alpha + \beta$. To prove Lemma 4.3, we need the following claim proved in [15]. For completeness, we include the short proof after finishing the proof of Lemma 4.3.

Claim 4.6. *Given independent random variables $A^*, B^* \in [T]$ in a probability space p', if A^* is γ_1-close to uniform, and B^* is γ_2-close to uniform, then*

$$p'(A^* = B^*) \geq \frac{1 - \gamma_1 - \gamma_2}{T}\,.$$

When c is good, Pinsker's inequality (Proposition 2.6) implies that conditioned on c, A is α-close to uniform and B is β-close to uniform. Then, since $A - C - B$, using Claim 4.6 (with $A^* = A$ and $B^* = B$ in the probability space p' conditioned on c) implies that $p'(A = B \mid c) \geq (1-\alpha-\beta)/T$. Since $p'(A = B) = 1/T$, we have that for a good c,

$$p'(c \mid A = B) = \frac{p'(c) \cdot p'(A = B \mid c)}{p'(A = B)} \geq (1-\alpha-\beta) \cdot p'(c)\,. \tag{4.8}$$

For any event $\mathcal{Q}$, (4.8) implies that

$$\begin{aligned} p'(C \in \mathcal{Q}) - p'(C \in \mathcal{Q} \mid A = B) &\leq \sum_{c \in \mathcal{Q}, c \text{ bad}} p'(c) + \sum_{c \in \mathcal{Q}, c \text{ good}} (p'(c) - p'(c \mid A = B)) \\ &\leq p'(C \text{ is bad}) + \sum_c p'(c)(\alpha+\beta) \\ &\leq \alpha + \beta + \sum_c p(c)(\alpha+\beta) \leq 2\alpha + 2\beta\,, \end{aligned}$$

and since $|p'(C) - p'(C \mid A = B)| = \max_{\mathcal{Q}}(p'(C \in \mathcal{Q}) - p'(C \in \mathcal{Q} \mid A = B))$ we get the required bound bound on statistical distance.

Proof of Claim 4.6. For each i, let $p'(A^* = i) = 1/T + \alpha_i$ and $p'(B^* = i) = 1/T + \beta_i$. Then, $\sum_i \alpha_i = \sum_i \beta_i = 0$, and $\alpha_i, \beta_i \geq -1/T$. Using these facts,

$$\begin{aligned} p'(A^* = B^*) &= \sum_i \left(\frac{1}{T} + \alpha_i\right)\left(\frac{1}{T} + \beta_i\right) \\ &= \frac{1}{T} + \frac{\sum_i \alpha_i}{T} + \frac{\sum_i \beta_i}{T} + \sum_i \alpha_i \beta_i = \frac{1}{T} + \sum_i \alpha_i \beta_i\,. \end{aligned}$$

To get a lower bound on the above, we will only consider the negative terms in the summation:

$$p'(A^* = B^*) \geq \frac{1}{T} + \sum_{i : \alpha_i > 0, \beta_i < 0} \alpha_i \beta_i + \sum_{i : \alpha_i < 0, \beta_i > 0} \alpha_i \beta_i \geq \frac{1}{T} - \frac{1}{T} \sum_{i : \alpha_i > 0} \alpha_i - \frac{1}{T} \sum_{i : \beta_i > 0} \beta_i\,.$$

From Proposition 2.1, it follows that $\sum_{i : \alpha_i > 0} \alpha_i$ is the statistical distance between A^* and the uniform distribution on $[T]$ and likewise for B^*. So, we get,

$$p'(A^* = B^*) \geq \frac{1 - \gamma_1 - \gamma_2}{T}\,. \qquad \square$$

4.3 Fooling distributions and relative discrepancy

In this section, we compare our techniques to that of Ganor, Kol and Raz [15]. The key concept introduced in [15] to prove lower bounds is the notion of the *relative discrepancy*. Let $f(x, y)$ be a Boolean function and $q(X, Y)$ be a distribution such that $q(f = 0) = q(f = 1) = 1/2$. Then, f has (ε, δ) *relative discrepancy* under $q(X, Y)$ if there exists a distribution $u(X, Y)$ such that for every rectangle $S \times T$ in the input space,

$$\left.\begin{matrix} q(X \in S, Y \in T \mid f = 0) \\ q(X \in S, Y \in T \mid f = 1) \end{matrix}\right\} \geq (1-\varepsilon) u(X \in S, Y \in T) \ \text{ if } u(X \in S, Y \in T) \geq \delta. \tag{4.9}$$

Ganor, Kol and Raz [15] proved that if f has $(\varepsilon = 1/3, \delta)$ relative discrepancy under $q(X,Y)$, then f has communication complexity $\Omega(\log 1/\delta)$. They then prove a lower bound on the relative discrepancy of the bursting noise function to give a communication lower bound. Note that the relative discrepancy technique individually argues about each rectangle that has large measure under the distribution $u(X,Y)$. Ganor, Kol and Raz [15] prove a lemma (Lemma 12 in [15]) that is very similar to our Lemma 4.1, however, they argue about each rectangle with a large enough measure under $u(X,Y)$.

In contrast, the fooling distribution technique allows us to argue about the distribution *on average* as opposed to arguing about each rectangle individually and this is where our proof becomes more simpler and more intuitive. Lemma 4.1 is an average-case version of the lemma proved in [15]. Together with our generalization of Shearer's Lemma (Lemma 4.2) and Lemma 4.3 we can argue about the messages on average. Note that our fooling distribution $p(X,F,Y,G)$ is analogous to the distribution $u(X,Y)$ in the definition of relative discrepancy.

Next we show some connections between relative discrepancy and fooling distributions. In fact, we show that low relative discrepancy implies the existence of a fooling distribution. The converse does not appear to be true.

Claim 4.7 (Relative Discrepancy implies Fooling Distribution)**.** *If f has relative discrepancy $(\varepsilon, 2^{-2\ell})$ then there exists a distribution $u(X,Y)$ such that if $M \in \{0,1\}^\ell$ denotes the messages of a deterministic protocol, then*

$$q(M \mid f=0) \overset{\gamma}{\approx} u(M) \overset{\gamma}{\approx} q(M \mid f=1)$$

where $\gamma = 2^{-\ell} + \varepsilon$.

Proof. Let $u(X,Y)$ be the distribution that satisfies (4.9) with $\delta = 2^{-2\ell}$. We will show that $u(M) \approx q(M \mid f=0)$. The proof for $u(M) \approx q(M \mid f=1)$ is similar. Define $\mathcal{B} = \{m \mid u(m) < 2^{-2\ell}\}$ and note that $u(M \in \mathcal{B}) < 2^{\ell} 2^{-2\ell} = 2^{-\ell}$. Also observe that when $m \notin B$, then by (4.9) and Proposition 2.13, $u(m) - q(m \mid f=0) \le \varepsilon u(m)$. Now for any event $\mathcal{Q}$, we have

$$\begin{aligned} u(M \in \mathcal{Q}) - q(M \in \mathcal{Q} \mid f=0) &\le \sum_{m \in \mathcal{Q} \cap \mathcal{B}} u(m) + \sum_{m \in \mathcal{Q} \cap \bar{\mathcal{B}}} (u(m) - q(m \mid f=0)) \\ &\le u(\mathcal{B}) + \varepsilon u(\bar{\mathcal{B}}) < 2^{-\ell} + \varepsilon\,. \end{aligned}$$

Hence $|u(M) - q(M \mid f=0)| = \max_{\mathcal{Q}}(u(M \in \mathcal{Q}) - q(M \in \mathcal{Q} \mid f=0)) < 2^{-\ell} + \varepsilon$. ☐

The above lemma gives another reason as to why our proof is simpler than the one given in [15]—we are proving a weaker statement that still implies a communication lower bound. We mention that in [15], a relaxed notion of relative discrepancy, called the *adaptive* relative discrepancy is also defined. Adaptive relative discrepancy works with a partition of the input space into rectangles as opposed to working with each rectangle individually, and an upper bound on this measure also suffices to prove a communication lower bound. Using arguments similar to Claim 4.7 one can prove that the existence of a fooling distribution implies that the adaptive relative discrepancy is small. Hence, the lower bounds given by the fooling distribution technique are sandwiched between the lower bounds that can be obtained by the relative discrepancy and the adaptive relative discrepancy methods.

We remark that the results of Fontes et al. [12] do not rule out the possibility that one might be able to separate information and communication complexity in the non-distributional setting (see Section 1.1) by using either of the two techniques, fooling distributions or adaptive relative discrepancy. In fact, Fontes et al. [12] show that the adaptive relative discrepancy is equal to the worst-case distributional communication complexity of the function up to polynomial factors, but we do not know if the same holds for fooling distributions.

5 Information upper bound

Let us recall the trivial protocol and the low-information protocol mentioned in Section 3. In the trivial protocol Alice and Bob send each other $X(z_{\leq i}), Y(z_{\leq i})$ in each step i, until they have computed z. The parties can compute $F(z)+G(z) \bmod 2$ with two more bits of communication. This protocol reveals at least $\Omega(\log n)$ bits of information as both parties learn the value of J with high probability. But note that the z computed in the above manner is consistent and the input distribution $q(X,F,Y,G)$ ensures that $F(z)+G(z) \bmod 2$ is the same for *every* consistent z, so it is enough for the parties to find a consistent z to complete the goal.

For the low-information protocol, the parties send each other the values $X(z_{\leq i}), Y(z_{\leq i})$ with probability $1-\varepsilon$ and send a uniformly random value otherwise. The parties abort the protocol if they see r rounds where the messages they sent were not the same. The distribution on inputs ensures that they will sample a consistent z with probability $1-O(\varepsilon)$. When the parties sample a consistent z, the messages sent are almost always sampled from a distribution that the receiving party knows, while if they sample a z that is not consistent, the protocol aborts shortly after the inconsistency.

The purpose of the noise is to prevent revealing a lot of information about J to both parties. When the z that the parties sample is consistent they only know that the value of J is among the locations where the values they have exchanged disagree. Since in this case there are εn disagreements in expectation, we intuitively expect that the entropy of J should still be large at the end which means that the information revealed about the value of J is small. In case the z sampled is not consistent (which happens with probability $O(\varepsilon)$), the abort condition prevents the parties from revealing too much information. By choosing the parameter ε carefully we can ensure that the total amount of information revealed is small.

In Figure 2 below we describe the low-information protocol which is parameterized by ε.

Lemma 5.1. *The protocol in Figure 2 outputs the correct answer with probability at least* $1-4\varepsilon$ *on inputs drawn from the distribution* $q(X,F,Y,G)$.

Proof. Under the input distribution $q(X,F,Y,G)$, whether a sample z with $|z| \geq J+1$ is consistent or not is determined solely by the $(J+1)^{\text{st}}$ step. So, the probability that the parties sample a z with $|z| \geq J+1$ and z being inconsistent is at most 2ε. Given that this event does not happen, the probability that the parties abort in a particular step is at most $(2\varepsilon)^{r-1}$, since to abort one of them must choose to send uniform values in at least $r-1$ steps where their inputs are equal. By the union bound, the probability that the parties abort in any step is at most $n(2\varepsilon)^{r-1}$. If neither of these bad events happens, the protocol computes the correct answer. Thus the probability of making an error is at most $2\varepsilon + n(2\varepsilon)^{r-1} \leq 4\varepsilon$, by the choice of r. □

The following theorem proves that the information cost of the protocol is small. We directly argue about the average information revealed about the messages, as compared to the proof of [15] who use the

Input: Alice is given (x, f), Bob is given (y, g). Both know a parameter $\varepsilon \in (0,1)$.
Output: $f(z) + g(z) \bmod 2$ for some consistent z.

Set z to be the empty string;
Set $r = \lceil \frac{\log n}{\log(1/2\varepsilon)} + 2 \rceil$;
for $i = 1, 2, \ldots, n$ **do**

Alice sets $a_i = \begin{cases} x(z_{<i}) & \text{with probability } 1-\varepsilon \\ \text{uniform element of } [k] & \text{with probability } \varepsilon \end{cases}$,

Bob sets $b_i = \begin{cases} y(z_{<i}) & \text{with probability } 1-\varepsilon \\ \text{uniform element of } [k] & \text{with probability } \varepsilon \end{cases}$;

Send $m_i = a_i, b_i$ to each other;
Set $w \in [k]$ so that $w = a_i + b_i \bmod k$, and append w to the string z;
if *$i \geq r$ and $a_{i'} \neq b_{i'}$ for all i' with $i - r + 1 \leq i' \leq i$* ; `// if` a_i `and` b_i `disagree in each of the last` r `rounds`
then
Terminate the protocol;
end

end
Send $f(z), g(z)$;

Figure 2: Protocol π_ε .

notion of the *tree divergence cost* to argue about the same.

Theorem 5.2. *If M denotes the messages in the protocol of Figure 2,*

$$\left.\begin{array}{l} \mathbf{I}_q(M : XF \mid YG) \\ \mathbf{I}_q(M : YG \mid XF) \end{array}\right\} \leq 2\log(k/\varepsilon) \cdot \left(1 + 16\varepsilon \cdot (\log n + 3) \cdot 2^{(2\log n)/(k\log(1/2\varepsilon))}\right).$$

Setting $\varepsilon = 1/\log n$ gives Theorem 1.1. To prove Theorem 5.2, we bound $\mathbf{I}_q(M : XF \mid YG)$. The second term is bounded in the same way.

Let $M_i \in [k] \times [k]$ be the messages exchanged in the i^{th} round for $i \in [n]$ and $M_{n+1} \in \{0,1\} \times \{0,1\}$ be the messages exchanged in the $(n+1)^{\text{th}}$ round. Then, by the chain rule, we can write

$$\mathbf{I}_q(M : XF \mid YG) = \mathop{\mathbb{E}}_{q(xfyg)} \left[\frac{q(M \mid xfyg)}{q(M \mid yg)} \right] = \sum_{i=1}^{|m|} \mathop{\mathbb{E}}_{q(mxfyg)} \left[\frac{q(M_i \mid m_{<i}xfyg)}{q(M_i \mid m_{<i}yg)} \right]. \tag{5.1}$$

To bound (5.1), we apply Proposition 2.11 which allows us to replace the distribution $q(M_i \mid m_{<i}yg)$ in the divergence expression above with a different distribution at the expense of increasing the divergence. Define distribution $u(MYG) = q(MYG \mid XF = YG)$. It will be simpler to analyze the divergence with respect to this distribution. Using Proposition 2.11, we then have

$$(5.1) \leq \sum_{i=1}^{|m|} \mathop{\mathbb{E}}_{q(mxfyg)} \left[\frac{q(M_i \mid m_{<i}xfyg)}{u(M_i \mid m_{<i}yg)} \right].$$

Next, we prove the following claim, which bounds the contribution to the divergence of each possible message.

Claim 5.3.

$$\frac{q(M_i \mid m_{<i}xfyg)}{u(M_i \mid m_{<i}yg)} \begin{cases} = 0 & \text{if } i \leq n \text{ and } x(z_{<i}) = y(z_{<i}), \\ \leq \log(k/\varepsilon) & \text{if } i \leq n \text{ and } x(z_{<i}) \neq y(z_{<i}), \\ = 1 & \text{if } i = n+1. \end{cases}$$

Before proving Claim 5.3, we show how to use it to bound the information. First note that the above claim intuitively says that information is revealed by M_i only when $X(Z_{<i}) \neq Y(Z_{<i})$. Recall that with probability $1 - O(\varepsilon)$, the parties sample a consistent z and hence, information is revealed only in one step of the protocol (and also one bit at the end when exchanging values of $F(z)$ and $G(z)$). However, in case the parties do not sample a consistent z, there might be a lot of messages M_i such that $X(Z_{<i}) \neq Y(Z_{<i})$.

To bound the information, next we show that on average the number of such messages is small since the parties will abort if they see a lot of disagreements. Towards this end, define $Q_i = 1$ if $|M| \geq i$ and $X(Z_{<i}) \neq Y(Z_{<i})$, and 0 otherwise. Claim 5.3 implies that

$$\mathbf{I}_q(M : FX \mid YG) \leq 1 + \log(k/\varepsilon) \cdot \mathop{\mathbb{E}}_{q} \left[\sum_{i=1}^{n} Q_i \right],$$

so it only remains to bound $\mathbb{E}_q \left[\sum_{i=1}^{n} Q_i \right]$ which is the number of messages where information is revealed.

Claim 5.4.

$$\mathop{\mathbb{E}}_{q} \left[\sum_{i=1}^{n} Q_i \right] \leq 1 + \frac{2r\varepsilon}{(1-1/k)^r} \leq 1 + 16\varepsilon \cdot (\log n + 3) \cdot 2^{(2\log n)/(k \log(1/2\varepsilon))}.$$

Claims 5.3 and 5.4 complete the proof of Theorem 5.2. We prove them next.

Proof of Claim 5.3. For any $i \in [n+1]$, let us write $M_i = A_i, B_i$ where A_i denotes Alice's message and B_i denotes Bob's message. Note that, when $i = n+1$, A_i and B_i are bits, and otherwise they are numbers in $[k]$. By the chain rule, for every $i \in [n+1]$,

$$\frac{q(M_i \mid m_{<i}xfyg)}{u(M_i \mid m_{<i}yg)} = \frac{q(A_i \mid m_{<i}xfyg)}{u(A_i \mid m_{<i}yg)} + \mathbb{E}_{q(a_i \mid m_{<i}xfyg)} \left[\frac{q(B_i \mid m_{<i}xfyga_i)}{u(B_i \mid m_{<i}yga_i)} \right]. \tag{5.2}$$

By the definition of the protocol in Figure 2, for every $i \in [n+1]$, B_i is determined by $YGM_{<i}$. This is because for $i \in [n]$, B_i is either a uniform random value or $Y(z_{<i})$ for some $z_{<i}$ determined by $M_{<i}$ and when $i = n+1$, $B_i = G(z)$ for some z determined by $M_{\leq n}$. It follows that for $i \in [n+1]$, $q(B_i \mid m_{<i}xfyga_i) = q(B_i \mid m_{<i}yg)$. Similarly, by the definition of the distribution $u(MYG)$, we have

$u(B_i \mid m_{<i}xfyga_i) = u(B_i \mid m_{<i}yg) = q(B_i \mid m_{<i}yg)$ for every $i \in [n+1]$. Hence, the second term in (5.2) is zero since the distributions are the same. So, for $i \in [n+1]$ we get that

$$\frac{q(M_i \mid m_{<i}xfyg)}{u(M_i \mid m_{<i}yg)} = \frac{q(A_i \mid m_{<i}xfyg)}{u(A_i \mid m_{<i}yg)}. \tag{5.3}$$

When $i = n+1$, using (5.3), a direct calculation shows

$$\frac{q(M_{n+1} \mid m_{\leq n}xfyg)}{u(M_{n+1} \mid m_{\leq n}yg)} = 1 \cdot \log(1/2) = 1\,.$$

Let us consider the case when $i \in [n]$ now. If $x(z_{<i}) = y(z_{<i})$, the definition of the protocol given in Figure 2, together with (5.3) ensures that $u(A_i \mid m_{<i}yg) = q(A_i \mid m_{<i}xfyg)$, proving the first bound. On the other hand if $x(z_{<i}) \neq y(z_{<i})$, then $q(A_i = a_i \mid m_{<i}xfyg) = u(A_i = a_i \mid m_{<i}yg)$, except when $a_i = x(z_{<i})$ or $a_i = y(z_{<i})$ (since otherwise the probability is ε/k in each case). Thus the divergence can be bounded by the contribution of these two values. We have that $q(A_i = x(z_{<i}) \mid m_{<i}xfyg) = (1-\varepsilon) + (\varepsilon/k)$ where the first and second terms account respectively for the cases when Alice sends the correct value and when Alice sends a random value. On the other hand, we have

$$u(A_i = x(z_{<i}) \mid m_{<i}yg) = q(A_i = x(z_{<i}) \mid m_{<i}yg, X = y, F = g) = \varepsilon/k\,.$$

Similarly, we get that

$$q(A_i = y(z_{<i}) \mid m_{<i}xfyg) = \varepsilon/k \quad \text{and} \quad u(A_i = y(z_{<i}) \mid m_{<i}yg) = 1 - \varepsilon + (\varepsilon/k)\,.$$

Therefore, using (5.3) we can bound

$$\begin{aligned}\frac{q(M_i \mid m_{<i}xfyg)}{u(M_i \mid m_{<i}yg)} &= (1-\varepsilon+(\varepsilon/k))\log\frac{1-\varepsilon+(\varepsilon/k)}{\varepsilon/k} + (\varepsilon/k)\log\frac{\varepsilon/k}{1-\varepsilon+\varepsilon/k} \\ &\leq \log\frac{1-\varepsilon+(\varepsilon/k)}{\varepsilon/k} \leq \log(k/\varepsilon)\,,\end{aligned}$$

as required. □

Proof of Claim 5.4. Let T be such that M_T is the last message sent in the protocol. Let $\mathcal{W}$ be the event that $T > J$ and Z sampled by the protocol is not consistent. Since $q(\mathcal{W}) \leq 2\varepsilon$ (if the parties do not send $X(Z_{\leq J})$ and $Y(Z_{\leq J})$ at the $(J+1)^{\text{st}}$ step), and $\mathbb{E}_q\left[\sum_{i=1}^n Q_i \mid \neg\mathcal{W}\right] \leq 1$ (at most one message where information is revealed when Z is consistent), we have

$$\mathop{\mathbb{E}}_q\left[\sum_{i=1}^n Q_i\right] \leq 2\varepsilon \cdot \mathop{\mathbb{E}}_q\left[\sum_{i=1}^n Q_i \;\middle|\; \mathcal{W}\right] + 1.$$

If $T > J$, $\sum_{i=1}^n Q_i = \sum_{i=J+1}^T Q_i \leq T - J$, so we get $\mathbb{E}_q\left[\sum_{i=1}^n Q_i \mid \mathcal{W}\right] \leq \mathbb{E}_q\left[T - J \mid \mathcal{W}\right]$. Note that $T-J$ roughly behaves like a geometric random variable. To bound this expectation, let us bound the contribution when $T-J \le r$ and when $T-J > r$ separately. The probability that the protocol does not abort in r rounds conditioned on the event $\mathcal{W}$ is at most $(1-1/k)^r$, so we have

$$\begin{aligned}\mathop{\mathbb{E}}_q[T-J \mid \mathcal{W}] &\le q(T-J \le r \mid \mathcal{W})r + q(T-J > r \mid \mathcal{W})(r + \mathop{\mathbb{E}}_q[T-J \mid \mathcal{W}]) \\ &\le (1-1/k)^r r + (1-(1-1/k)^r)(r + \mathop{\mathbb{E}}_q[T-J \mid \mathcal{W}]) \\ \implies \mathop{\mathbb{E}}_q[T-J \mid \mathcal{W}] &\le \frac{r}{(1-1/k)^r},\end{aligned}$$

as required. The second inequality in the statement of the claim follows from the fact that $1/(1-1/k) \le 2^{2/k}$, for $k \ge 2$, and by the choice of r. $\square$

6 Acknowledgments

We thank Paul Beame, Yuval Filmus, Moni Naor, Sivaramakrishnan Ramamoorthy and Ran Raz for useful conversations and anonymous referees for careful reading and very helpful comments.

References

[1] ZIV BAR-YOSSEF, T. S. JAYRAM, RAVI KUMAR, AND D. SIVAKUMAR: An information statistics approach to data stream and communication complexity. *J. Comput. System Sci.*, 68(4):702–732, 2004. Conference version in FOCS'02. [doi:10.1016/j.jcss.2003.11.006] 2

[2] BOAZ BARAK, MARK BRAVERMAN, XI CHEN, AND ANUP RAO: How to compress interactive communication. *SIAM J. Comput.*, 42(3):1327–1363, 2013. Conference version in STOC'10. [doi:10.1137/100811969] 2

[3] MARK BRAVERMAN: Interactive information complexity. *SIAM J. Comput.*, 44(6):1698–1739, 2015. Conference version in STOC'12. [doi:10.1137/130938517] 2, 3, 4, 11

[4] MARK BRAVERMAN AND ANKIT GARG: Public vs private coin in bounded-round information. In *Proc. 41st Internat. Colloq. on Automata, Languages and Programming (ICALP'14)*, pp. 502–513. Springer, 2014. Full version ECCC TR13-130. [doi:10.1007/978-3-662-43948-7_42] 2

[5] MARK BRAVERMAN AND ANUP RAO: Information equals amortized communication. *IEEE Trans. Inform. Theory*, 60(10):6058–6069, 2014. Conference version in FOCS'11. [doi:10.1109/TIT.2014.2347282, arXiv:1106.3595] 2

[6] MARK BRAVERMAN, ANUP RAO, OMRI WEINSTEIN, AND AMIR YEHUDAYOFF: Direct products in communication complexity. In *Proc. 54th FOCS*, pp. 746–755. IEEE Comp. Soc. Press, 2013. [doi:10.1109/FOCS.2013.85] 2, 6, 7

[7] MARK BRAVERMAN AND OMRI WEINSTEIN: A discrepancy lower bound for information complexity. *Algorithmica*, 76(3):846–864, 2016. Conference version in APPROX'12. [doi:10.1007/s00453-015-0093-8, arXiv:1112.2000] 2, 4

[8] AMIT CHAKRABARTI, YAOYUN SHI, ANTHONY WIRTH, AND ANDREW CHI-CHIH YAO: Informational complexity and the direct sum problem for simultaneous message complexity. In *Proc. 42nd FOCS*, pp. 270–278. IEEE Comp. Soc. Press, 2001. [doi:10.1109/SFCS.2001.959901] 2

[9] FAN R. K. CHUNG, RONALD L. GRAHAM, PETER FRANKL, AND JAMES B. SHEARER: Some intersection theorems for ordered sets and graphs. *J. Combin. Theory Ser. A*, 43(1):23–37, 1986. [doi:10.1016/0097-3165(86)90019-1] 13, 14

[10] THOMAS M. COVER AND JOY A. THOMAS: *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006. 6

[11] URIEL FEIGE, PRABHAKAR RAGHAVAN, DAVID PELEG, AND ELI UPFAL: Computing with noisy information. *SIAM J. Comput.*, 23(5):1001–1018, 1994. Conference version in STOC'90. [doi:10.1137/S0097539791195877] 10

[12] LILA FONTES, RAHUL JAIN, IORDANIS KERENIDIS, SOPHIE LAPLANTE, MATHIEU LAURIÈRE, AND JÉRÉMIE ROLAND: Relative discrepancy does not separate information and communication complexity. *ACM Trans. Comput. Theory*, 9(1):4:1–4:15, 2016. Conference version in ICALP'15. [doi:10.1145/2967605] 4, 21, 22

[13] ANAT GANOR, GILLAT KOL, AND RAN RAZ: Exponential separation of information and communication. In *Proc. 55th FOCS*, pp. 176–185. IEEE Comp. Soc. Press, 2014. [doi:10.1109/FOCS.2014.27] 3

[14] ANAT GANOR, GILLAT KOL, AND RAN RAZ: Exponential separation of communication and external information. In *Proc. 48th STOC*, pp. 977–986. ACM Press, 2016. [doi:10.1145/2897518.2897535] 4

[15] ANAT GANOR, GILLAT KOL, AND RAN RAZ: Exponential separation of information and communication for Boolean functions. *J. ACM*, 63(5):46:1–46:31, 2016. Conference version in STOC'15. [doi:10.1145/2907939] 3, 4, 6, 7, 10, 14, 19, 20, 21, 22

[16] PRAHLADH HARSHA, RAHUL JAIN, DAVID A. MCALLESTER, AND JAIKUMAR RADHAKRISHNAN: The communication complexity of correlation. *IEEE Trans. Inform. Theory*, 56(1):438–449, 2010. Conference version in CCC'07. [doi:10.1109/TIT.2009.2034824] 2

[17] BALA KALYANASUNDARAM AND GEORG SCHNITGER: The probabilistic communication complexity of set intersection. *SIAM J. Discrete Math.*, 5(4):545–557, 1992. Conference version in 2nd Structure in Complexity Theory Conf., 1987. [doi:10.1137/0405044] 2

[18] IORDANIS KERENIDIS, SOPHIE LAPLANTE, VIRGINIE LERAYS, JÉRÉMIE ROLAND, AND DAVID XIAO: Lower bounds on information complexity via zero-communication protocols and applications. *SIAM J. Comput.*, 44(5):1550–1572, 2015. Conference version in FOCS'12. [doi:10.1137/130928273, arXiv:1204.1505] 2

[19] GILLAT KOL: Interactive compression for product distributions. In *Proc. 48th STOC*, pp. 987–998. ACM Press, 2016. [doi:10.1145/2897518.2897537] 2

[20] EYAL KUSHILEVITZ AND NOAM NISAN: *Communication Complexity*. Cambridge University Press, 1997. 8

[21] SIVARAMAKRISHNAN NATARAJAN RAMAMOORTHY AND ANUP RAO: How to compress asymmetric communication. In *Proc. 30th Conf. on Computational Complexity (CCC'15)*, pp. 102–123. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2015. [doi:10.4230/LIPIcs.CCC.2015.102] 2

[22] SIVARAMAKRISHNAN NATARAJAN RAMAMOORTHY AND MAKRAND SINHA: On the communication complexity of greater-than. In *Proc. 53rd Ann. Allerton Conf. on Communication, Control, and Computing (Allerton'15)*, pp. 442–444. IEEE Comp. Soc. Press, 2015. [doi:10.1109/ALLERTON.2015.7447037] 4

[23] JAIKUMAR RADHAKRISHNAN: Entropy and counting. In *Computational Mathematics, Modelling and Algorithms*, pp. 146–168. Narosa Publishers, 2003. Available from author's home page. 13, 14

[24] ANUP RAO AND MAKRAND SINHA: Simplified separation of information and communication. *Electron. Colloq. on Comput. Complexity (ECCC)*, 2015. [ECCC:TR15-057]

[25] ALEXANDER A. RAZBOROV: On the distributional complexity of disjointness. *Theoret. Comput. Sci.*, 106(2):385–390, 1992. Conference version in ICALP'90. [doi:10.1016/0304-3975(92)90260-M] 2

[26] CLAUDE E. SHANNON: A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. [doi:10.1002/j.1538-7305.1948.tb01338.x] 2

[27] ALEXANDER A. SHERSTOV: Compressing interactive communication under product distributions. *SIAM J. Comput.*, 47(2):367–419, 2018. Conference version in FOCS'16. [doi:10.1137/16M109380X] 2

[28] EMANUELE VIOLA: The communication complexity of addition. *Combinatorica*, 35(6):703–747, 2015. Conference version in SODA'13. [doi:10.1007/s00493-014-3078-3] 4

AUTHORS

Anup Rao
Associate professor
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA
anuprao@cs.washington.edu
homes.cs.washington.edu/~anuprao/

Makrand Sinha
Ph. D. student
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA
makrand@cs.washington.edu
homes.cs.washington.edu/~makrand/

ABOUT THE AUTHORS

ANUP RAO is an associate professor of computer science at the Allen School of Computer Science & Engineering, University of Washington. He is interested in mathematics and theoretical computer science. When he is not exercising his brain on these topics, he likes to exercise his body rock climbing, skiing, going on alpine adventures, dancing and playing pickup basketball with undergraduates who become reluctant to foul him once they figure out that he is a professor. He loves to travel, and loves art of all kinds. Anup feels fortunate to have been advised by David Zuckerman.

MAKRAND SINHA is a graduate student at the Allen School of Computer Science & Engineering, University of Washington where he is advised by Anup Rao. He is truly interested in all facets of the human experience, but currently his efforts are focused towards understanding communication complexity and linear programs. When not scribbling mathematics on a piece of paper, he can be found enjoying fine coffee, literature, theater and a multitude of physical activities.